Am J Cancer Res 2023;13(11):5493-5503
www.ajcr.us /ISSN:2156-6976/ajcr0152118

Original Article

Deep learning-based diagnosis of lung cancer using a nationwide respiratory cytology image set: improving accuracy and inter-observer variability

Taehee Kim[1*], Hyun Chang[2*], Binna Kim[3], Jaeho Yang[4], Dongjun Koo[5], Jeongwon Lee[6], Ji Wouk Chang[7], Gisu Hwang[8], Gyungyub Gong[4], Nam Hoon Cho[9], Chong Woo Yoo[10], Ju-Yeon Pyo[9], Yosep Chong[3]

[1]Division of Pulmonary, Allergy and Critical Care Medicine, Department of Internal Medicine, Hallym University Kangnam Sacred Heart Hospital, Hallym University College of Medicine, Seoul, South Korea; [2]Division of Medical Oncology and Hematology, Department of Internal Medicine, International St. Mary's Hospital, Catholic Kwandong University College of Medicine, Incheon, South Korea; [3]Department of Pathology, Uijeongbu St. Mary's Hospital, College of Medicine, The Catholic University of Korea, Gyeonggi-do, South Korea; [4]Department of Pathology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, South Korea; [5]Interdisciplinary Program in Bioengineering, College of Engineering, Seoul National University, Seoul, South Korea; [6]Osong Medical Innovation Foundation, Chungcheongbuk-do, South Korea; [7]Seegene Medical Foundation Inc., Seoul, South Korea; [8]AI Team, DeepNoid Inc., Seoul, South Korea; [9]Department of Pathology, Yonsei University College of Medicine, Seoul, South Korea; [10]Department of Pathology, National Cancer Center, Ilsan, Gyeonggi-do, South Korea. [*]Equal contributors.

Received July 9, 2023; Accepted September 18, 2023; Epub November 15, 2023; Published November 30, 2023

Abstract: Deep learning (DL)-based image analysis has recently seen widespread application in digital pathology. Recent studies utilizing DL in cytopathology have shown promising results, however, the development of DL models for respiratory specimens is limited. In this study, we designed a DL model to improve lung cancer diagnosis accuracy using cytological images from the respiratory tract. This retrospective, multicenter study used digital cytology images of respiratory specimens from a quality-controlled national dataset collected from over 200 institutions. The image processing involves generating extended z-stack images to reduce the phase difference of cell clusters, color normalizing, and cropping image patches to 256 × 256 pixels. The accuracy of diagnosing lung cancer in humans from image patches before and after receiving AI assistance was compared. 30,590 image patches (1,273 whole slide images [WSIs]) were divided into 27,362 (1,146 WSIs) for training, 2,928 (126 WSIs) for validation, and 1,272 (1,272 WSIs) for testing. The Densenet121 model, which showed the best performance among six convolutional neural network models, was used for analysis. The results of sensitivity, specificity, and accuracy were 95.9%, 98.2%, and 96.9% respectively, outperforming the average of three experienced pathologists. The accuracy of pathologists after receiving AI assistance improved from 82.9% to 95.9%, and the inter-rater agreement of Fleiss' Kappa value was improved from 0.553 to 0.908. In conclusion, this study demonstrated that a DL model was effective in diagnosing lung cancer in respiratory cytology. By increasing diagnostic accuracy and reducing inter-observer variability, AI has the potential to enhance the diagnostic capabilities of pathologists.

Keywords: Lung cancer, cytopathology, digital pathology, deep learning, convolutional neural network

Introduction

Lung cancer is the leading cause of cancer-related death worldwide [1]. Because approximately 70% of newly diagnosed lung cancer is found in the unresectable advanced stage, early detection and diagnosis are the most important issues to improve its prognosis [2]. Several small studies have investigated the use of chest X-rays and sputum cytology for lung cancer screening, but these tests were found to have limited sensitivity and specificity for detecting lung cancer in its early stages [3-6]. The results of the National Lung Screening Trial (NLST), the first large-scale clinical trial, showed that low-dose computed tomography (LDCT) scans were significantly more effective than chest X-rays at detecting early-stage lung cancer and reducing lung cancer-related mortality by 20% [7]. Based on the results of NLST, sev-

eral countries, including the United States, Canada, Japan, and South Korea, have adopted LDCT screening for lung cancer. However, current evidence shows that LDCT screening for lung cancer is associated with significant harms, including overdiagnosis, false positives, consequences of invasive follow-up procedures, procedure related complications, and relatively higher test cost [8].

Cytological examination has the advantages of inexpensive, rapid, and minimally invasive procedures for cancer screening. However, compared to histopathological examination, cytopathologic diagnosis has the disadvantages of not only low diagnostic accuracy but also labor-intensiveness, time-consumption, and inter-observer variation in interpretation [9, 10]. The diagnostic performance of cytology for lung cancer is relatively low: sensitivity of exfoliative sputum cytology (0.49-0.71) and of abrasive cytology obtained from bronchoscopy (0.43-0.59) [11].

The application of artificial intelligence (AI), especially deep learning (DL) techniques, to image analysis may offer a promising alternative to conventional lung cancer screening tools by augmenting the accuracy of cytopathology diagnostic performance. Recently, deep learning techniques inspired by the mechanisms of vision have been widely applied to image classification, object detection, and prediction by utilizing multilayer neural networks, namely convolutional neural networks (CNNs) [12]. CNNs have also been widely applied in the medical imaging field, especially digital pathology, suggesting potential for use in clinical pathology [13-16]. In comparison, there have been few studies on the application of AI in the field of digital cytopathology [17, 18]. Several studies on digital cytopathology image analysis of lung cancer have reported the potential usefulness of DL-based classification models (binary classification [malignant vs. benign disease] and lung cancer subtypes classification) [19-22]. However, these results are limited by relatively low diagnostic accuracy, small sample size, and lack of pathologist involvement.

In this study, we aimed to design a DL algorithm to improve the accuracy of lung cancer diagnosis by pathologists from cytological images of respiratory specimens using a quality-controlled, nationally representative dataset.

Materials and methods

Study design

In this retrospective, multicenter study, we collected digital cytopathology images of respiratory specimens from sputum and bronchial washing, corresponding to clinical information. The study was approved by the Institutional Review Board of the Catholic University of Korea, College of Medicine (UC21SNSI0064), the Institutional Review Board of the Yonsei University College of Medicine (4-2021-0569), and the Institutional Review Board of the National Cancer Center (NCC2021-0145). The review boards waived the requirement for written informed consent because of the retrospective study design, and data were collected and anonymized according to confidentiality guidelines. **Figure 1** shows a schematic overview of the method and workflow proposed in this paper.

Datasets

The study utilized respiratory cytology specimens obtained from sputum and bronchial washing in "The OPEN AI Dataset PROJECT" for training, validation, and testing (Supplementary Figure 1). This project is part of the AI learning data construction project conducted by the Korea Intelligence Information Society Agency since 2017 and aims to provide AI learning data, software, and computing resources essential for AI technology and service development. This is an open-source public dataset available at "AI-hub (https://aihub.or.kr/)" and we utilized respiratory cytology specimens obtained from sputum and bronchial washing fluid collected between 2021 and 2022 among the non-gynecologic cytopathology image datasets from body fluids.

This dataset comprises digitalized cytopathology slides obtained from the Quality Control Committee of the Korean Society of Cytopathology. The cytology image dataset of respiratory specimens consisted of whole slide images (WSI) from more than 200 universities, general hospitals, and laboratory centers in Korea, collected for use in quality control programs (20%) and three tertiary hospitals (Catholic University Medical Uijeongbu St. Mary's Hospital, Yonsei University Severance Hospital, and National Cancer Center) and

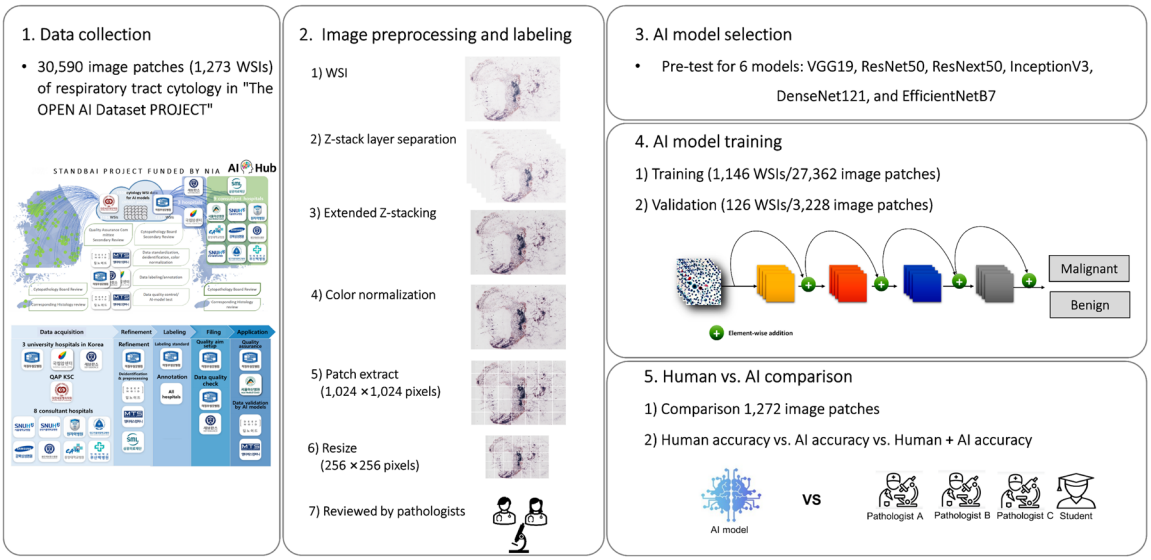


Figure 1. Flow chart of the study design. Abbreviations: AI, artificial intelligence; WSI, whole slide image.

eight major hospitals nationwide (80%). The WSIs were obtained by scanning the slides at focal planes at 40× magnification using Pannoramic 250 Flash III (3DHistech, Hungary), AT2 (Leica Biosystems, Germany), and NanoZoomer S360 (Hamamatsu, Japan) scanners. All cytopathology specimens were reviewed by pathologists, who compared cytopathology images with histopathology diagnoses to select matched pathologic confirmation and typical textbook cases from quality control programs (20%) and collected daily practice cases from three tertiary hospitals and eight major hospitals nationwide (80%). Specimens of cygology collected from each institution were initially reviewed for corresponding histology by board-certified pathologists of the respective institutions. Subsequently, they underwent a secomd review by the Quality Assurance Committee. Following the scanning of the cytology specimens on glass slides, the scanned images underwent refinement and validation. During the labeling process, image quality was reexamined to account for factors such as discoloration, artifacts, air bubbles, and out-of-focus areas, and images with low image quality or no corresponding tissue diagnosis were excluded (Supplementary Figure 1).

Labeling was initially performed on WSIs during scanning according to each institution's diagnostic classification criteria and was primarily based on the corresponding histopathology diagnosis on cytology slides from the same participant. The basic characteristics of the corresponding cytology image datasets, including the participant's age, gender, histological type, cytological preparation, and staining method, were collected from medical records of participating hospitals.

Image preprocessing

To reduce the phase difference of cell clusters while digitizing cell slides, extended Z-stack images were generated through separating and integrating of Z-stack image layers, as presented in **Figure 2**. Color normalization was performed on the integrated images prior to patch extraction. The resulting images were subsequently cropped to dimensions of 1,024 × 1,024 pixels and resized to 256 × 256 pixels to serve as input for model training. Each patch image was evaluated by experienced pathologists to ensure accurate diagnosis. 30,590 image patches were obtained from 1,273 respiratory tract WSIs.

For image patches extracted from WSIs, normal slides were assigned benign disease class annotations for all image patches. Lung cancer slides were reclassified as normal or malignant by two or more experts (cytotechnologists and cytopathologists) after reviewing the extracted image patches. During this process, data that did not meet the quality standards were excluded. In addition, for image patches extracted from lung cancer WSIs, if there was disagree-

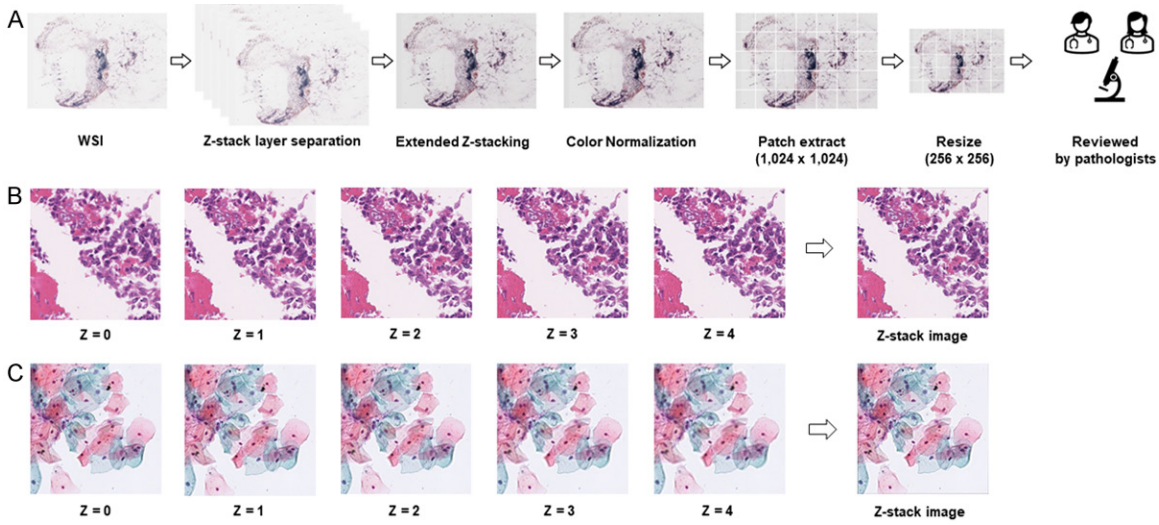


Figure 2. Image preparation process. Image preprocessing and patch extraction from whole slide image and resizing on digitalized respiratory cytology images (A). Generation of extended Z-stacking images in H&E stain (B), and PAP stain (C). Abbreviations: WSI, whole slide image; H&E, hematoxylin and eosin; PAP, Papanicolaou.

ment between two experts during the review process, those image patches were excluded from the image dataset.

Deep learning model training

To identify the most optimal model for our study, we evaluated and compared the accuracy of six representative convolutional neural network (CNN) models: VGG19, ResNet50, ResNext50, InceptionV3, DenseNet121, and EfficientNetB7. For training and validation of the model, we used 10,069 image patches extracted from 716 malignant WSIs and 20,521 image patches extracted from 557 benign WSIs, which were divided into a hematoxylin and eosin (H&E) staining set and papanicolaou (PAP) staining set according to the staining methodology.

Outcomes: comparison of performance between human and AI

We randomly selected 1,272 patches out of a total of 30,590 image patches and used them to compare the diagnostic accuracy of humans and AI. The diagnostic performance of the AI was evaluated by three experienced pathologists and one trained non-medical student. Furthermore, the diagnostic accuracy of the pathologists was reevaluated after they referred to the results obtained from AI. All pathologists and students conducted their reevaluation after a washout period of two weeks, and in both evaluation (before and after the AI model diagnostic reference) were conducted without knowledge of the ground truth.

Statistical analysis

To evaluate the AI model performance, we used Python (version 3.11.2) to calculate several metrics. We calculated the numbers of true positives, false negatives, true negatives, and false positives at a classification threshold of 0.5 using the sklearn.metrics.confusion_matrix function. From these values, we calculated the sensitivity, specificity, accuracy, positive predictive value (PPV), and negative predictive value (NPV) using standard formulas. We also calculated 95% confidence intervals for these metrics using exact binomial confidence limits. To assess inter-rater reliability among multiple raters, we calculated Fleiss' kappa coefficient using R (version 4.1.2). This was done using the fleiss.kappa function from the package. The input to this function is a matrix whose rows each represent an item being rated and columns each represent a rater. The entries in the matrix are the ratings assigned by each rater to each item. The fleiss.kappa function returns the Fleiss' kappa coefficient, which measures the degree of agreement among the raters. The t-test was conducted using the SPSS statistical software package, and the significance level was set at $P < 0.05$. A paired t-test was used to compare the mean diagnostic accuracy before and after referencing the AI results.

Table 1. Train and validation dataset split for AI model training

	H&E stain			PAP stain		
	Benign	Malignancy	Total	Benign	Malignancy	Total
Train	4,786 (135)	651 (92)	5,437 (227)	13,902 (367)	8,047 (552)	21,925 (919)
Validation	533 (15)	36 (10)	569 (25)	1,300 (40)	1,335 (61)	2,659 (101)
Total	5,319 (150)	687 (102)	6,006 (252)	15,202 (407)	9,382 (613)	24,584 (1,020)

Data are presented as number of image patches (whole slide images). H&E, hematoxylin and eosin; PAP, Papanicolaou.

Table 2. Confusion matrix and diagnostic performance of AI model using DenseNet121 network

		Ground truth									
		H&E stain					PAP stain				
		Benign	Malignancy	Total			Benign	Malignancy	Total		
Prediction	Benign	529	4	533	Sensitivity	0.889	1,227	61	1,288	Sensitivity	0.955
	Malignancy	4	32	36	Specificity	0.993	73	1,371	1,444	Specificity	0.944
	Total	533	36	569	Accuracy	0.986	1,300	1,432	2,732	Accuracy	0.950

H&E, hematoxylin and eosin; PAP, Papanicolaou.

Results

Participant information

The detailed clinical and pathological information used in this study is summarized in Supplementary Table 1. A total of 1,273 patients were included in the study. The mean age of the patients was 65.6 ± 12.6 years. Of the patients, 884 (69.4%) were male and 389 (30.6%) were female. The histopathologic diagnosis were lung malignancies in 716 (56.2%) and benign diseases in 557 (43.8%) patients. Cytology preparation was performed using conventional smear in 346 (27.2%) patients and liquid-based cytology in 927 (72.8%) patients. Staining was performed using H&E stain in 252 (19.8%) patients and PAP stain in 1,021 (80.2%) patients.

Training and validation data

For the H&E stain, the training dataset consisted of 4,786 image patches of benign cases and 651 image patches of malignant cases, for a total of 5,437 image patches (**Table 1**). The validation dataset consisted of 533 image patches of benign cases and 36 image patches of malignant cases, for a total of 569 image patches. Overall, there were 5,319 and 687 image patches of benign and malignant cases for the H&E stain, respectively. For the PAP stain, the training dataset consisted of 13,902 image patches of benign cases and 8,047 image patches of malignant cases, for a total of 21,925 image patches. The validation dataset consisted of 1,300 image patches of benign cases and 1,335 image patches of malignant cases, for a total of 2,659 image patches. Overall, there were 15,202 and 9,382 image patches of benign and malignant cases for the PAP stain, respectively.

Pre-test for AI model selection

The preliminary analysis was conducted to test 6 different AI models for their diagnostic performance in terms of sensitivity, specificity, and accuracy for each model and stain type (Supplementary Table 2). For H&E stain, DenseNet121 had the highest sensitivity of 0.889 and specificity of 0.993, resulting in an accuracy of 0.986. For PAP stain, DenseNet121 had the highest sensitivity of 0.956 and specificity of 0.944, resulting in an accuracy of 0.950. These results suggest that DenseNet121 has the potential to perform well in the detection of malignant cells using both H&E and PAP stains. Therefore, DenseNet121 was selected for further analysis.

AI model training

Table 2 summarizes the confusion matrix and diagnostic performance of an AI model using the DenseNet121 network for both H&E and PAP stains. For H&E stain, the model had a sensitivity of 0.889 and specificity of 0.993 for the benign and malignant categories, respectively, resulting in an accuracy of 0.986. For PAP stain, the model had a sensitivity of 0.955 and specificity of 0.944 for the benign and malignant cat-

Table 3. Comparison of diagnostic performance between AI and human

	Sensitivity	Specificity	Accuracy	*p*-value*	Fleiss' Kappa
Student	71.5%	86.5%	77.4%	$P < 0.05$	
Student + AI	94.4%	94.4%	94.4%		
Pathologist A	76.5%	93.0%	83.7%	$P < 0.05$	
Pathologist A + AI	95.4%	98.7%	96.9%		
Pathologist B	77.1%	90.8%	83.1%	$P < 0.05$	
Pathologist B + AI	95.4%	98.2%	96.6%		
Pathologist C	95.8%	64.3%	82.0%	$P < 0.05$	
Pathologist C + AI	95.0%	93.4%	94.3%		
Average Pathologists	83.1%	82.7%	82.9%	$P < 0.05$	0.553
Pathologists + AI	95.2%	96.8%	95.9%		0.908
AI	95.9%	98.2%	96.9%		

AI, artificial intelligence. *Paired t-test.

egories, respectively, resulting in an accuracy of 0.950.

Comparison between AI and human using randomly selected image patches

To compare the performance of three pathologists, one student, and AI in diagnosing lung cancer from respiratory cytology images, we randomly selected 1,272 image patches from the dataset. The patch images included 557 benign lesion and 715 lung cancer cytopathological images. The results show that AI assistance improved the diagnostic performance of both the student and pathologists (**Table 3** and **Figure 3A**). The diagnostic performance of the AI system was high, with a sensitivity of 95.9%, specificity of 98.2%, and accuracy of 96.9%. Compared to humans, the AI system showed higher diagnostic performance.

Without AI assistance, the sensitivity, specificity, and accuracy of the student were 71.5%, 86.5%, and 77.4%, respectively. With AI assistance, these values increased to 94.4%, 94.4%, and 94.4%, respectively. The average sensitivity, specificity and accuracy of pathologists are 83.1%, 82.7% and 82.9%, respectively. When pathologists work with AI, their sensitivity, specificity, and accuracy increase significantly to 95.2%, 96.8% and 95.9% respectively. Additionally, The Fleiss' Kappa scores increased from 0.553 before AI assistance to 0.908 after AI assistance, improving inter-examiner agreement in three pathologists. We also confirmed in the correlation plot that the diagnostic agreement of pathologists and students improved before (**Figure 3B**) and after (**Figure 3C**) referring to the AI reading results. Examples of correctly and incorrectly diagnosed image patches by the AI are shown in Supplementary Figure 2, and examples from pathologists are shown in Supplementary Figure 3.

Discussion

In this study, we used a dataset of respiratory cytologic images that maximized the generalizability for AI models by considering different genders, ages, diagnoses, staining methods, scanner types, and geographical distributions, collected from over 200 hospitals in South Korea. We found that a DL model showed promising results in distinguishing lung cancer cells from benign cells in respiratory cytology and could improve the diagnosis of pathologists by balancing sensitivity and specificity and reducing inter-observer variation.

This demonstrates sufficient reliability in terms of performance compared to previous studies that were conducted at a single institution with relatively small sample sizes (Supplementary Table 3). Teramoto et al. [19] constructed a DL model for lung cancer subtype classification of 298 cytology image patches from fine needle aspiration (FNA) and bronchoscopy into three subtypes of small cell carcinoma (SCLC), adenocarcinoma (ADC), and squamous cell carcinoma (SqCC), which showed an accuracy of 71.1%. In the following years, the same research team sequentially reported a CNN model to classify benign and malignant disease with an accuracy of 79.2% on 621 image patches [20] and an accuracy of 85.3% on 793 image patches [21]. In addition, Gonzalez et al. trained the Inception V3 CNN model to distinguish SCLC from large cell neuroendocrine carcino-

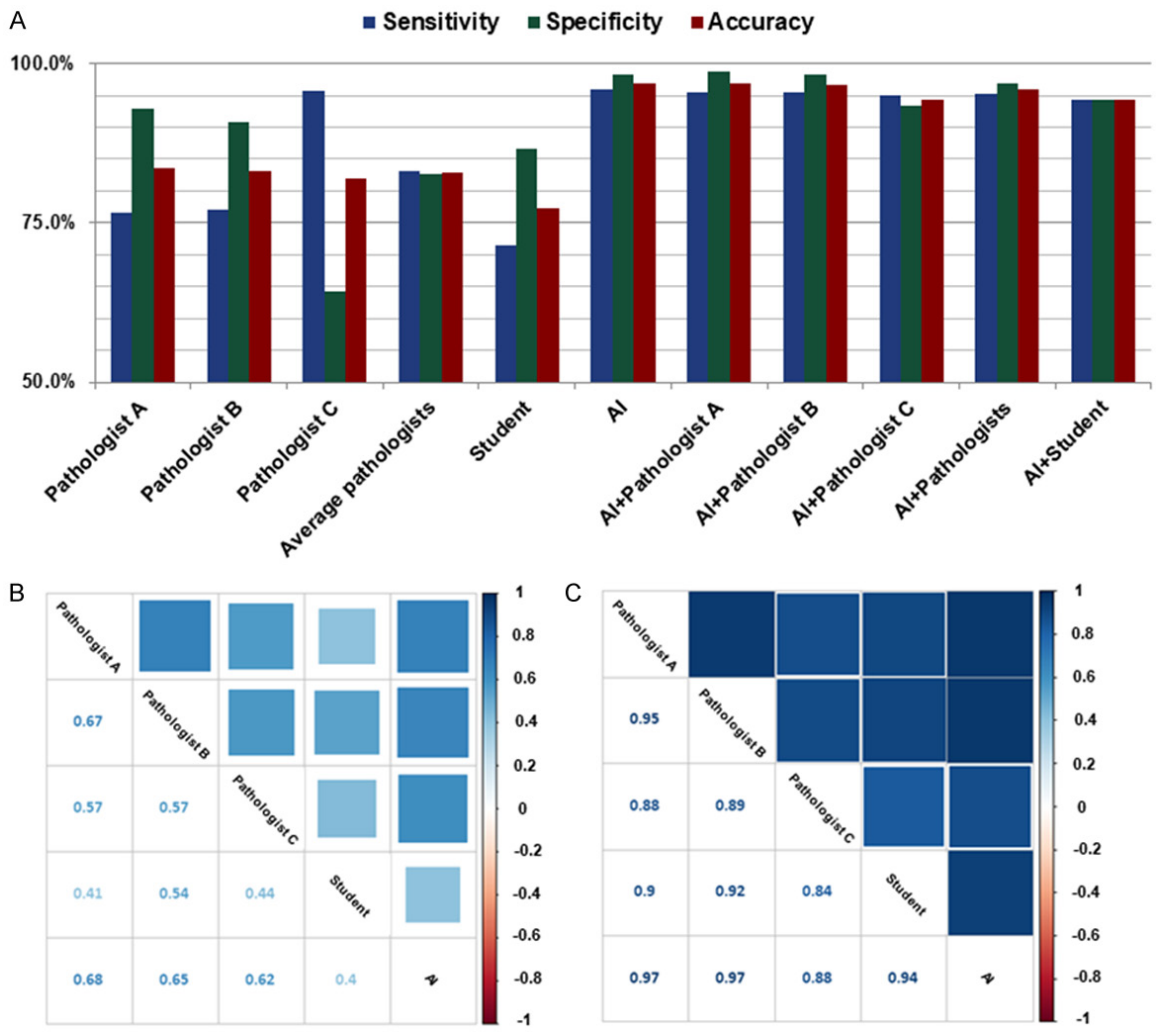


Figure 3. Comparison of diagnostic performance between human and AI model in respiratory tract cytologic images. Comparison of diagnosis results of 3 pathologists, 1 non-medical major student, and AI (A). Correlation plot of diagnoses before (B) and after referencing AI results (C). Abbreviation: AI, artificial intelligence.

ma using 114 WSIs of FNA samples [22]. The reported performance was a sensitivity of 100%, specificity of 87.5%, and AUC of 100% on a Diff-Quik stain dataset, 100%, 85.7%, and 100% on a PAP stain set, and 100%, 87.5%, and 87.5% on a H&E stain set, respectively. In a 2021 study, Teramoto et al. used multiple instance learning (MIL) and several CNN architectures to achieve a 91.6% accuracy in binary classification with 322 image patches [23]. Meanwhile, Tsukamoto et al. [24] classified three lung cancer types using 298 image patches from various architectures: AlexNet (73.7% accuracy), GoogLeNet/InceptionV3 (66.8%), VGG-16 (76.8%), and ResNet-50 (74.0%). Although the performance of these studies was relatively good, the small dataset size from a single institute cannot guarantee the generalizability that can be applied to daily practice. In contrast to previous studies, our study utilized a larger dataset (30,590 image patches from 1,274 WSIs) from various institutions, which was large enough to compensate for dataset imbalance, prevent algorithmic bias and ensure the reliability of the DL diagnostic algorithm.

One of the important lessons from the results of this study is that AI will be a potential augmentation tool for human pathologists by enhancing and adjusting the imbalanced performance of each observer. As shown in **Table 3**, pathologists A and B exhibited similar diagnostic tendencies with relatively low sensitivity (76.5% and 77.1%, respectively) and relatively

high specificity (93% and 90.8%, respectively). In contrast, pathologist C had relatively high sensitivity (95.8%) but relatively low specificity (64.3%). This discrepancy can result in false positives or false negatives, leading to a decrease in diagnostic accuracy. In this study, the AI-assisted diagnosis achieved a sensitivity of 95.2%, which was 12% higher than the average pathologist sensitivity, and a specificity of 96.8%, which was 14% higher than the average pathologist specificity. We can see that the low sensitivity of pathologist A and B, and the low specificity of pathologist C were greatly improved after referring the AI diagnosis and the level of agreement on the diagnosis between pathologists was also greatly improved. This indicates that the AI system could assist in increasing diagnostic performance of pathologists by enhancing the weaker part of each pathologist.

Another interesting finding was the results of the student (**Table 3**). The participant student was a student researcher majored in the computer science without biological background or knowledge of cytopathology. We taught him the cytologic findings and characteristics of benign and malignant cells for few hours and let him diagnose the image patches as benign or malignant. The accuracy of the student was fairly good for the first timer as 77.4%. However, more interesting thing was his accuracy after the referring the AI results increased greatly to 94.4% which is higher than the accuracy of average pathologists. This suggest AI can be helpful as a second opinion for the trainees and it could be used as an aid to reduce the time and cost to educate cytopathologists in the field of cytopathology, which typically requires significant time and field experience to become an expert. Many countries, including Korea, are facing a shortage of cytologists and cytopathologists due to global ageing and increasing cytologic exams that will cause them at risk of misdiagnosis due to the increased workload from cancer patients. Applying AI aids to cytopathologic diagnosis can be the potential solution to the fundamental problems of labor-intensity, time-consumption, and low-accuracy.

In comparing the diagnostic performance of human and AI, we reviewed the correctly diagnosed positive and negative cases by AI as well as the misdiagnosed positive and negative cases (Supplementary Figure 2). False positive and negative cases (Supplementary Figure 2B and 2C) were all correctly diagnosed by human pathologists. Reactive alveolar macrophages ingesting foreign particles and debris often presents a wide spectrum of morphologic findings that are sometimes very challenging, even to expert pathologists. Also, reactive bronchial epithelial cells are commonly found due to secondary infection and inflammation from various causes as well as mechanical obstruction of cancer mass. In this situation, pathologists make a diagnosis by comprehensively collecting information both from background and individual cells and clusters. However, it is not always easy to collect sufficient samples containing tumor cells and background due to sampling difficulty. When either a typical background or individual cell features of malignancy is missing, it can be very challenging for pathologists to make a correct diagnosis. In these examples, the cytologic findings including nuclear atypia and structural abnormality of the cell clusters were generally understandable for pathologists in true positive and true negative cases (Supplementary Figure 2A and 2D). However, some of the cytologic findings of false positive and false negative cases were not easily understandable representing the black box-like nature of AI interpretation. There were cases that the AI model over- or underdiagnosed while all the human pathologists correctly diagnosed (Supplementary Figure 3A and 3B) and all the human pathologists over- or underdiagnosed while only AI model correctly answered (Supplementary Figure 3C and 3D). The most cases that the AI model overdiagnosed were from artifacts such as mucin materials and pyknotic cells in the bloody background that mimic dyskeratotic squamous cells (Supplementary Figure 3A). On the other hand, most false negative cases that the AI models misdiagnosed but all the pathologists correctly answered showed the relatively obvious cytologic findings of malignancy such as malignant glandular cells or squamous cells that made us question about the performance of the AI model showing the current limitation of black box-like nature of the AI models (Supplementary Figure 3B). However, there were also the cases that only AI answered correctly while all the human pathologists misdiagnosed. The false positive cases by the human pathologists

showed the reactive bronchial cells, dyskeratotic squamous cells, or reactive macrophages with severe cytologic atypia (Supplementary Figure 3C). The false negative cases by the human pathologists were malignant dyskeratotic cells with less obvious cytologic atypia or low nuclear cytoplasmic ratio or small malignant blue round cells from small cell carcinoma with less obvious cytologic atypia (Supplementary Figure 3D). These findings can provide important insights into how AI can be utilized to improve diagnostic accuracy and consistency in the pathologists' interpretation. Some of the adenocarcinoma cells can be underdiagnosed as reactive bronchial cells or immature metaplasia and dyskeratotic squamous cells, one of the important diagnostic clues for squamous cell carcinoma, can be easily considered as dyskeratotic cells from severe inflammation and infectious condition.

One of the most promising applications of AI in lung cancer diagnosis will be subtyping of lung cancers on histologic or cytologic images and predicting driver mutations based on the morphologic findings as a screening test before confirmative molecular testing. In 2022, Yang et al. [25] developed a CNN model for subtyping of lung carcinoma biopsy and showed a promising results on the histologic images. Two studies by Teramoto et al. that were mentioned earlier also showed a promising results of subtyping of lung cancer from cytology images although the diagnostic accuracy is still relatively low and the dataset size is small to be applied in the clinical field and the sample was FNAs but not respiratory tract samples [19, 20, 22]. In addition to these subtyping AI models, AI models that can predict mutation such as EGFR, KRAS, ALK can be next important technology for expanding the impact of targeted therapy. Recently, Ren et al. [26] introduced a DL model to predict targeted gene alterations in lung cancer from pleural effusion cell block WSIs. The model was trained on ten genes related to targeted therapy and four genetic mutation statuses (ALK fusion, KRAS mutation, EGFR mutation, and no alterations group). The genetic mutation prediction results of the DL model were reasonable as AUC of 0.869 for ALK fusion, 0.804 for KRAS mutation, 0.644 for EGFR mutation, and 0.774 for no alterations. However, relatively small-sized and unevenly distributed datasets limited the ability to generalize the AI's diagnostic performance (e.g., 23 fusions and 335 wild-types for ALK, 215 mutations and 143 wild-types for EGFR, etc.). As this study has demonstrated the feasibility of molecular diagnosis at the cytology image level, further research will follow to prove the utility of DL for predicting molecular diagnosis.

The strength of our study is that, compared to previous studies, excellent performance has been demonstrated with high generalizability in terms of both dataset quality and DL diagnostic accuracy, as we removed biases in data collection and implemented quality control measures for all populations. AI diagnostic algorithms are generally designed to identify patterns in data rather than understand the underlying biological mechanisms of diseases. Therefore, poorly managed big data can have a negative impact on performance. Some limitations of this study should be acknowledged. First, lung cancer prediction models at the WSI level have yet to be fully explored. While pathologists generally make diagnoses by assessing WSIs rather than image patches, it is important to develop an algorithm that can ensure both WSI-level and single image patch-level accuracy to improve the accuracy and reliability of AI-based cytopathology techniques. Second, although we used a large-scale respiratory cytology image dataset representative of Korea, additional external validation is required to demonstrate the generalizability of the AI model. Third, it is necessary to further validate the accuracy of the AI model with data from different races and countries to demonstrate the scalability of the AI model.

Conclusion

The DL based diagnoses have shown promising results in distinguishing between lung cancer and benign cells in respiratory tract cytology, potentially assisting pathologists to balance sensitivity and specificity and reducing interobserver variability.

Acknowledgements

I appreciate Mr. June Lee for participating the data curation. This research was supported by Hallym University Medical Center Research Fund, a National Research Foundation of Korea (NRF) grant funded by the Korean government

(MSIT) (2021R1A2C2013630), and The Catholic University of Korea Uijeongbu St. Mary's Hospital Clinical Research Laboratory Foundation made in the program year of 2023.

Disclosure of conflict of interest

None.

Abbreviations

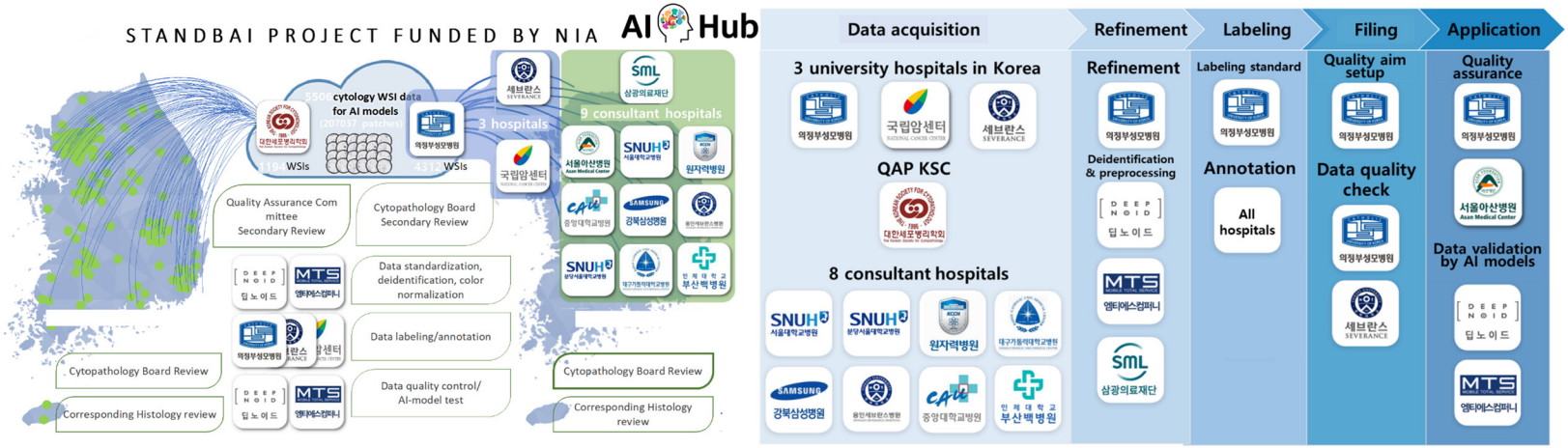
DL, deep learning; WSI, whole slide images; NLST, national lung screening trial; LDCT, low-dose computed tomography; AI, artificial intelligence; H&E, hematoxylin and eosin; PAP, papanicolaou; CNN, convolutional neural network; ROC, receiver operating characteristic; AUC, area under the ROC curve; PPV, positive predictive value; NPV, negative predictive value; FNA, fine needle aspiration; SCLC, small cell lung carcinoma; ADC, adenocarcinoma; SqCC, squamous cell carcinoma; EGFR, epidermal growth factor receptor; KRAS, Kirsten rat sarcoma viral oncogene; ALK, anaplastic lymphoma kinase.

Address correspondence to: Dr. Yosep Chong, Department of Pathology, Uijeongbu St. Mary's Hospital, College of Medicine, The Catholic University of Korea, #271, Cheonbo-ro, Uijeongbu 11765, Gyeonggi-do, South Korea. Tel: +82-031-820-3160; Fax: +82-031-820-3877; ORCID: 0000-0001-8615-3064; E-mail: ychong@catholic.ac.kr

References

[1] Siegel RL, Miller KD, Fuchs HE and Jemal A. Cancer statistics, 2022. CA Cancer J Clin 2022; 72: 7-33.
[2] Travis WD. Pathology of lung cancer. Clin Chest Med 2011; 32: 669-692.
[3] Dales LG, Friedman GD and Collen MF. Evaluating periodic multiphasic health checkups: a controlled trial. J Chronic Dis 1979; 32: 385-404.
[4] Flehinger BJ and Kimmel M. The natural history of lung cancer in a periodically screened population. Biometrics 1987; 43: 127-144.
[5] Kubik A and Haerting J. Survival and mortality in a randomized study of lung cancer detection. Neoplasma 1990; 37: 467-475.
[6] Marcus PM, Bergstralh EJ, Zweig MH, Harris A, Offord KP and Fontana RS. Extended lung cancer incidence follow-up in the Mayo Lung Project and overdiagnosis. J Natl Cancer Inst 2006; 98: 748-756.
[7] National Lung Screening Trial Research Team, Aberle DR, Adams AM, Berg CD, Black WC, Clapp JD, Fagerstrom RM, Gareen IF, Gatsonis C, Marcus PM and Sicks JD. Reduced lung-cancer mortality with low-dose computed tomographic screening. N Engl J Med 2011; 365: 395-409.
[8] Usman Ali M, Miller J, Peirson L, Fitzpatrick-Lewis D, Kenny M, Sherifali D and Raina P. Screening for lung cancer: a systematic review and meta-analysis. Prev Med 2016; 89: 301-314.
[9] Chong Y, Ji SJ, Kang CS and Lee EJ. Can liquid-based preparation substitute for conventional smear in thyroid fine-needle aspiration? A systematic review based on meta-analysis. Endocr Connect 2017; 6: 817-829.
[10] Al-Abbadi MA. Basics of cytology. Avicenna J Med 2011; 1: 18-28.
[11] Schreiber G and McCrory DC. Performance characteristics of different modalities for diagnosis of suspected lung cancer: summary of published evidence. Chest 2003; 123 Suppl: 115S-128S.
[12] LeCun Y, Boser B, Denker J, Henderson D, Howard R, Hubbard W and Jackel L. Handwritten digit recognition with a back-propagation network. Adv Neural Inf Process Syst 1989; 2.
[13] Niazi MKK, Parwani AV and Gurcan MN. Digital pathology and artificial intelligence. Lancet Oncol 2019; 20: e253-e261.
[14] Pallua JD, Brunner A, Zelger B, Schirmer M and Haybaeck J. The future of pathology is digital. Pathol Res Pract 2020; 216: 153040.
[15] Bera K, Schalper KA, Rimm DL, Velcheti V and Madabhushi A. Artificial intelligence in digital pathology-new tools for diagnosis and precision oncology. Nat Rev Clin Oncol 2019; 16: 703-715.
[16] Alam MR, Abdul-Ghafar J, Yim K, Thakur N, Lee SH, Jang HJ, Jung CK and Chong Y. Recent applications of artificial intelligence from histopathologic image-based prediction of microsatellite instability in solid cancers: a systematic review. Cancers (Basel) 2022; 14: 2590.
[17] Thakur N, Alam MR, Abdul-Ghafar J and Chong Y. Recent application of artificial intelligence in non-gynecological cancer cytopathology: a systematic review. Cancers (Basel) 2022; 14: 3529.
[18] Alrafiah AR. Application and performance of artificial intelligence technology in cytopathology. Acta Histochem 2022; 124: 151890.
[19] Teramoto A, Tsukamoto T, Kiriyama Y and Fujita H. Automated classification of lung cancer types from cytological images using deep convolutional neural networks. Biomed Res Int 2017; 2017: 4067832.
[20] Teramoto A, Yamada A, Kiriyama Y, Tsukamoto T, Yan K, Zhang L, Imaizumi K, Saito K and

Fujita H. Automated classification of benign and malignant cells from lung cytological images using deep convolutional neural network. Inform Med Unlocked 2019; 16: 100205.
[21] Teramoto A, Tsukamoto T, Yamada A, Kiriyama Y, Imaizumi K, Saito K and Fujita H. Deep learning approach to classification of lung cytological images: two-step training using actual and synthesized images by progressive growing of generative adversarial networks. PLoS One 2020; 15: e0229951.
[22] Gonzalez D, Dietz RL and Pantanowitz L. Feasibility of a deep learning algorithm to distinguish large cell neuroendocrine from small cell lung carcinoma in cytology specimens. Cytopathology 2020; 31: 426-431.
[23] Teramoto A, Kiriyama Y, Tsukamoto T, Sakurai E, Michiba A, Imaizumi K, Saito K and Fujita H. Weakly supervised learning for classification of lung cytological images using attention-based multiple instance learning. Sci Rep 2021; 11: 20317.
[24] Tsukamoto T, Teramoto A, Yamada A, Kiriyama Y, Sakurai E, Michiba A, Imaizumi K and Fujita H. Comparison of fine-tuned deep convolutional neural networks for the automated classification of lung cancer cytology images with integration of additional classifiers. Asian Pac J Cancer Prev 2022; 23: 1315-1324.
[25] Yang JW, Song DH, An HJ and Seo SB. Classification of subtypes including LCNEC in lung cancer biopsy slides using convolutional neural network from scratch. Sci Rep 2022; 12: 1830.
[26] Ren W, Zhu Y, Wang Q, Jin H, Guo Y and Lin D. Deep learning-based classification and targeted gene alteration prediction from pleural effusion cell block whole-slide images. Cancers (Basel) 2023; 15: 752.



Supplementary Figure 1. Overview of the open AI Dataset PROJECT.

Supplementary Table 1. Baseline characteristics of the enrolled patients

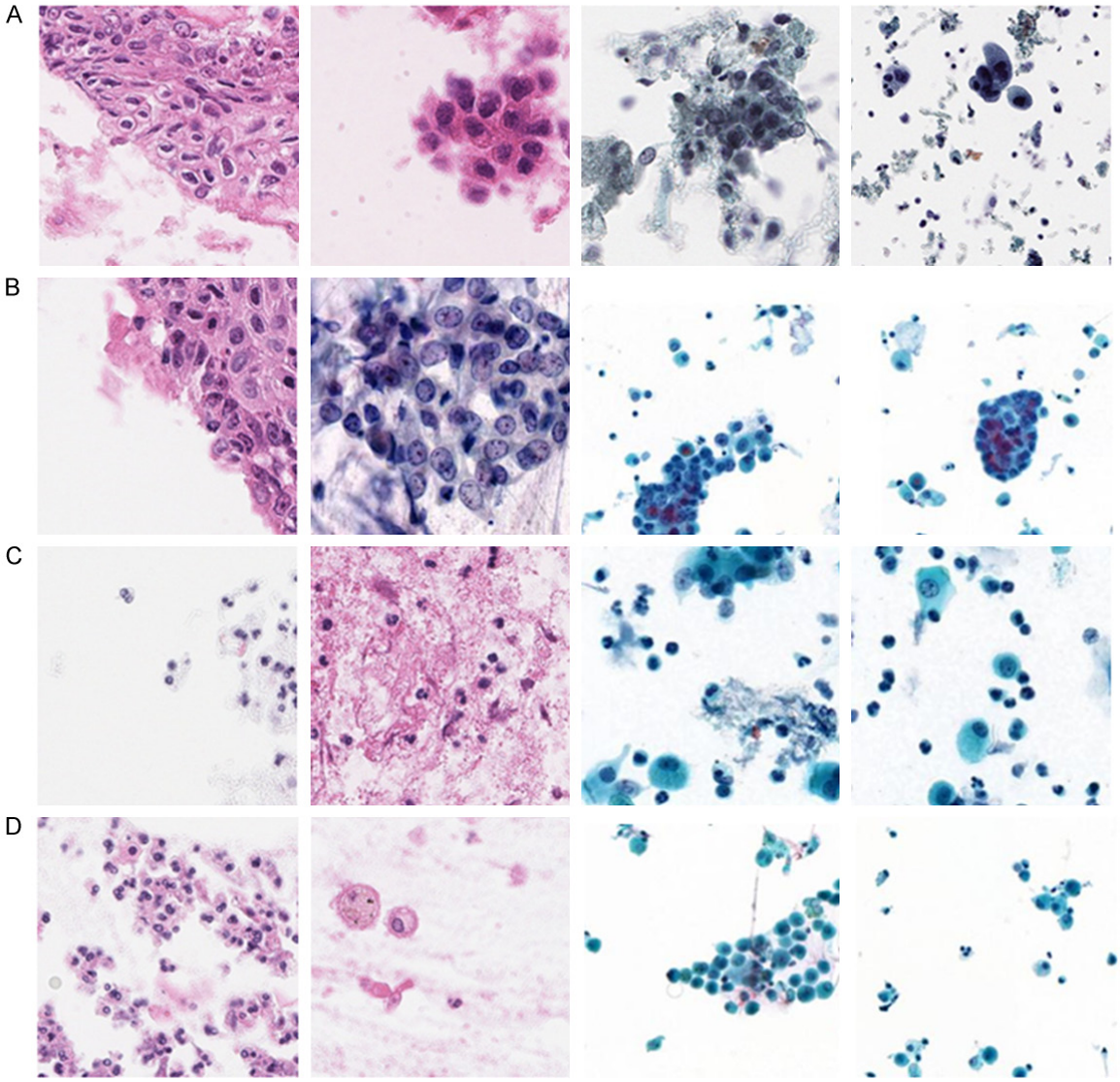
	Total (N = 1,273)
Age, years	65.6 ± 12.6
Gender	
Male	884 (69.4)
Female	389 (30.6)
Histopathologic diagnosis	
Malignancy	716 (56.2)
Benign lesions	557 (43.8)
Cytology preparation	
Conventional smear	346 (27.2)
Liquid based cytology	927 (72.8)
Stain method	
H&E stain	252 (19.8)
PAP stain	1,021 (80.2)

Data are presented as number (percentage) or means ± standard deviations. H&E, hematoxylin and eosin; PAP, Papanicolaou.

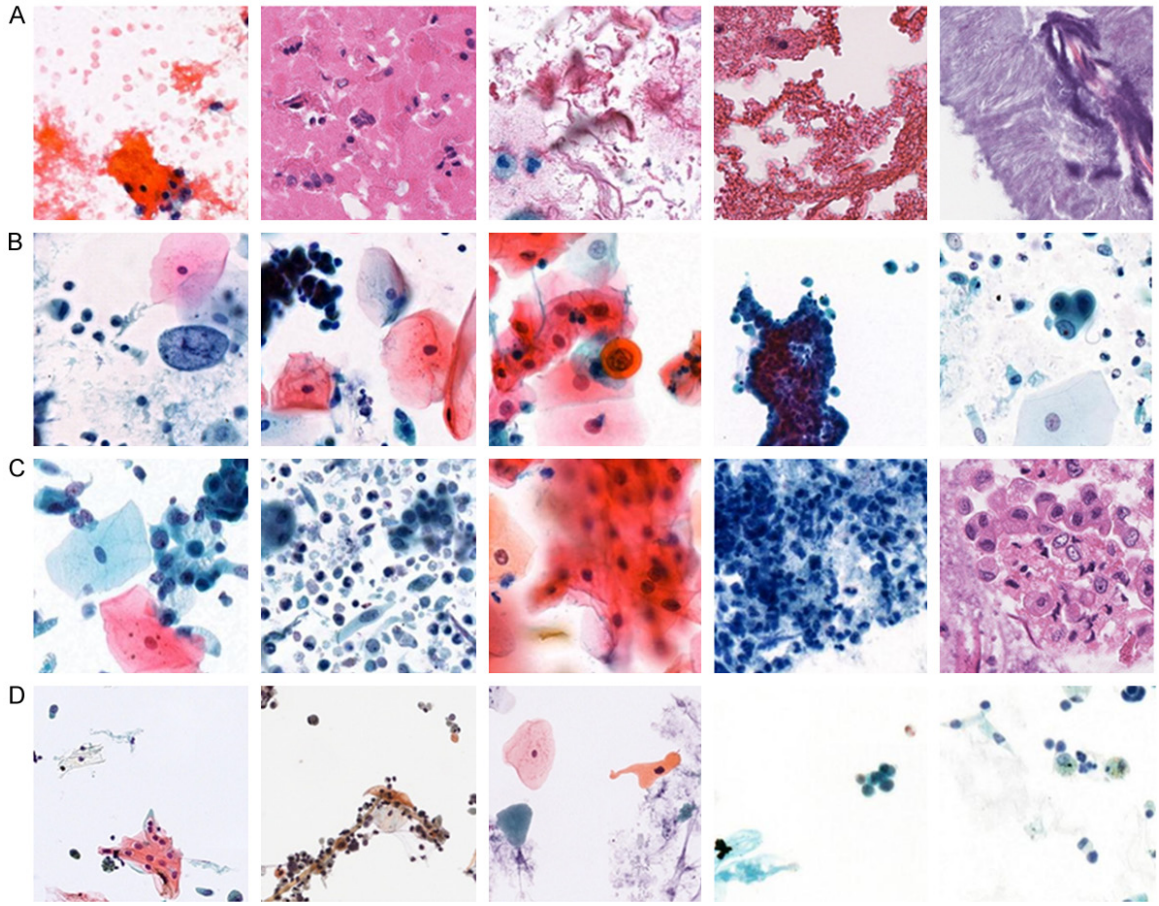
Supplementary Table 2. Diagnostic performance in preliminary test results of AI models

	H&E stain			PAP stain		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
VGG19	0.000	1.000	0.937	0.914	0.879	0.897
ResNet50	0.806	0.987	0.975	0.932	0.933	0.932
ResNext50	0.976	0.889	0.970	0.393	0.922	0.904
InceptionV3	0.861	0.994	0.986	0.935	0.914	0.925
DenseNet121	0.889	0.993	0.986	0.956	0.944	0.950
EfficientNetB7	0.694	0.993	0.974	0.893	0.952	0.921

AI, artificial intelligence; H&E, hematoxylin and eosin; PAP, Papanicolaou.



Supplementary Figure 2. Representative examples of AI prediction on cytology image patch from respiratory specimen. (A) True positive cases, (B) False negative cases, (C) False positive cases, and (D) True negative cases. Abbreviation: AI, artificial intelligence.



Supplementary Figure 3. Representative examples of human misdiagnosis in cytology image patches from respiratory specimens. A. False positive cases that all human pathologists made a correct diagnosis, but the AI model produced an incorrect result. B. False negative cases that all human pathologists made a correct diagnosis, but the AI model produced an incorrect result. C. False positive cases that all human pathologists made an incorrect diagnosis, but the AI model produced a correct diagnosis. D. False negative cases that all human pathologists made an incorrect diagnosis, but the AI model produced a correct diagnosis. Abbreviation: AI, artificial intelligence.

Supplementary Table 3. Previous studies applying AI models to cytology images

Author, year	Classification	Dataset	Model	Diagnostic performance
Teramoto, 2017 [19]	3-class: (ADC/SqCC/SCLC)	298 image patches from 76 cases	Custom CNN architecture	Accuracy: 71.1%
Teramoto, 2019 [20]	Binary: (Benign/Malignant)	621 image patches from 46 cases	VGG16	Sensitivity: 89.3% Specificity: 83.3% Accuracy: 79.2%
Teramoto, 2020 [21]	Binary: (Benign/Malignant)	793 image patches from 60 cases	Combination of progressive growing GAN and VGG16 architecture	Sensitivity: 85.4% Specificity: 85.3% Accuracy: 85.3%
Gonzalez, 2020 [22]	Binary: (SCLC/LCNEC)	464,378 image patches from 40 cases	InceptionV3	For Diff-Quik Sensitivity: 100% Specificity: 87.5% AUC: 100% For PAP Sensitivity: 100% Specificity: 85.7% AUC: 100% For H&E Sensitivity: 100% Specificity: 87.5% AUC: 87.5%
Teramoto, 2021 [23]	Binary: (Benign/Malignant)	322 image patches from 322 cases	MIL and several CNN architectures as backbone	Accuracy: 91.6%
Tsukamoto, 2022 [24]	3-class: (ADC/SqCC/SCLC)	298 image patches from 55 cases	AlexNet GoogLeNet/InceptionV3 VGG-16 ResNet-50	AlexNet Accuracy: 73.7% GoogLeNet/InceptionV3 Accuracy: 66.8% VGG16 Accuracy: 76.8% ResNet50 Accuracy: 74.0%

AI, artificial intelligence; ADC, adenocarcinoma; SqCC, squamous cell carcinoma; SCLC, small cell lung cancer; LCNEC, large cell neuroendocrine carcinoma; WSI, whole slide image; FNA, fine needle aspiration; CNN, convolutional neural network; GAN, generative adversarial network; AUC, area under the receiver operating characteristic curve; PAP, Papanicolaou; H&E, hematoxylin and eosin; MIL, multiple instance learning.